Extended Experimental Procedures

Fish housing and husbandry

Fish were raised at 25°C in a centralized filtration water system at a density of up to 1 fish per 1.4L in 2.8L and 9L tanks. Fish were fed freshly hatched *Artemia* nauplii until 3 weeks of age and then dried bloodworm (*Chironomous sp.*) twice a day during the week and once a day during weekends. Adult fish spawned on a sand substrate in 2.8L and 9L tanks within the centralized filtration system. Dead fish were removed daily from the tanks, weighted, and stored in 50 mL of 100% ethanol. Embryos were collected on a weekly basis and plated on sterile dry peat moss until they were ready to hatch. Once ready to hatch, indicated by the presence of a distinct yellow iris in the eye and by continuous body twitching within the chorion, embryos were immersed in a 4^{o}C Yamamoto embryo solution (17 mM NaCl, 2.7 mM KCl, 2.5 mM CaCl2, 0.02 mM NaHCO3 pH 7.3) (Rembold et al., 2006) supplemented with peat moss extract and oxygen tablets. Hatched fry were placed in 0.2-gallon tanks at the density of 5 fry per tank and fed with brine shrimp nauplii.

Koeppen-Geiger analysis of climate

Data on annual precipitation and temperature were collected in 102 meteorological stations from http://en.climate-data.org/ and were used to compute a Koeppen-Geiger climate classification in southeastern Africa according to the classification by Peel (Peel et al., 2007).

Next-generation sequencing of the turquoise killifish genome

Genomic DNA was isolated from tissues (muscle or tail) of 9 African turquoise killifish individuals (8 males and one female) from the inbred reference strain GRZ. The predominance of male individuals allows a more comprehensive survey of the genome of this species, because males are heterogametic in this species (Kirschner et al., 2012; Valenzano et al., 2009). Ten libraries with varying insert sizes were constructed for Illumina sequencing as indicated in the table below from 9 independent GRZ fish (one fish was used to build two libraries). Genomic sequences for de novo assembly were generated on Illumina HiSeq2000 instruments (Beijing Genome Institute, University of Oregon, and Stanford Center for Genomics and Personalized Medicine). Paired-end libraries were obtained as 2x101bp raw reads, and mate-pair libraries were obtained as 2x50bp raw reads. All genomic sequencing libraries have been deposited to SRA (SRP041421).

Library	Insert size (bp)	Strategy	Total QC length (bp)	Coverage (X)[a]	Center	Sex
BGI170	170	Paired-end	24,461,613,765	12.23	BGI	M
BGI500	500	Paired-end	27,538,490,563	13.77	BGI	M
GRZ340	340	Paired-end	9,892,583,577	4.95	Stanford	M
GRZ540	540	Paired-end	3,856,857,526	1.93	Stanford	M
RADGP0	200	Paired-end	1,892,127,538	0.95	Oregon	F
RADAAP0	200	Paired-end	1,032,838,324	0.52	Oregon	M
GRZ300	300	Paired-end	22,661,204,008	11.33	Stanford	M
GRZ400	400	Paired-end	22,531,661,933	11.27	Stanford	M
BGI2K	2,000	Mate-pair	32,648,404,824	16.32	BGI	M
BGI5K	5,000	Mate-pair	12,742,324,364	6.37	BGI	M
		Total	159,258,106,422	79.63		

[a] coverage estimate based on a conservative genome size estimate of 2Gb (see below). Note that the total coverage estimate ranges from 72.4X (for the maximal genome estimate of 2.2Gb) to 122.5X (for the minimal genome size estimate of 1.3Gb).

Read quality filtering and trimming for de novo genome and transcriptome assembly

Sequencing reads in all Illumina libraries for de novo genome and transcriptome assembly were quality filtered and trimmed using the trim_galore software (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), with a Phred score threshold of 30 and a minimum remaining read length of 50bp in either read of the pair after trimming. Due to nucleotide composition biases at the beginning of sequencing reads, all reads were also further trimmed of their first four most 5' bases using fastx_trimmer (http://hannonlab.cshl.edu/fastx_toolkit).

Genome size estimate from Illumina sequencing libraries

Genome size was estimated using two independent methods. The first method was based on (Li et al., 2010). Briefly, 25-mer frequencies were counted using the Jellyfish software (Marcais and Kingsford, 2011) and the corresponding maximum k-mer frequency was graphically determined. Genome size was estimated using the count of base pairs in the used reads, the mean length of the quality-trimmed reads, and the maximum k-mer frequency. With this method, the turquoise killifish genome size was estimated to be 1.9-2.2Gb, taking into account three independent libraries (GRZ300, GRZ340 and GRZ400).

The second method used the preqc module in SGA (Simpson, 2014). This method also takes a k-mer frequency approach, but accounts for sequencing error rates, potential heterozygous sites, and effect of repeat sequences (Simpson, 2014). The preqc module (SGA v 0.10.13) was run on three independent libraries (BGI170, GRZ300 and GRZ340). The genome size was estimated to be 1.3-1.6Gb by this method. Thus, the computational estimate of the turquoise killifish genome size ranges from 1.3-2.2Gb. This is consistent with (Reichwald et al., 2009). Based on this range, the percentage of genome assembled is 83.1% to 49.1%, respectively. In the manuscript, we use a conservative genome size estimate of 2Gb (~54% of genome assembled, with ~80 fold coverage).

De novo genome assembly with SGA and SOAPdenovo

The overlapping 170bp Illumina paired-end library was preprocessed to obtain bona fide longer sequence fragments using FLASH (Magoc and Salzberg, 2011). The length distribution of output fragments was compatible with the library specifications. SGA v0.9.19-10 (string graph assembler) (Simpson and Durbin, 2012) was applied to the filtered pre-processed Illumina paired-end libraries to obtain a high-quality master assembly. We chose SGA as our main contig assembler because of its lower reported rate of missassemblies compared to other assemblers (Salzberg et al., 2012). Final assembly parameters in SGA were a correction k-mer of 51, an overlap length of 65bp in the string-graph, and a merging overlap length of 75bp to generate the contig assembly. Parameters to the sga-assemble function were set to "-d 0.2 -g 0.1 -r 10" to account for potential indels and SNPs in the paired-end (PE) libraries (constructed from 7 independent GRZ individuals, 6 males and 1 female), in the range recommended in the software instructions.

A second assembly was obtained using the de Bruijn graph assembly SOAPdenovo V1.05 (Luo et al., 2012), with an assembly k-mer of 81 and using the error-corrected reads from the sga pipeline, requiring support of 5X coverage and minimum assembled length of 200bp for contig reporting.

Scaffolding, GapFilling and assembly reconciliation

Scaffolding was performed on the SGA and SOAP assemblies using the SSPACE Basic v2.0 scaffolder (Boetzer et al., 2011). All paired-end and mate-pair librairies (10 libraries total) were inputted in hierarchical size order, requiring support of ≥ 5 independent pairs to create a scaffold link. Parameters were set to: no contig extension, a 0.7 maximum link ratio, only 1bp allowed gap for alignment, and reporting only contigs longer than 200bp. Gap-filling was conducted on the scaffolded assemblies using GapFiller v1.10 (Nadalin et al., 2012) inputting all paired-end and mate-pair libraries in hierarchical size order. Parameters were set to: a minimum overlap with the gap of ≥ 31bp, a minimum of 5 supporting reads, a

minimum of 15bp overlap to merge sequences of a closing gap, a 15bp trimming, a 50bp gap-close difference, a base ratio of 0.7, all over 10 iterations.

Because of different biases associated with each assembly algorithms (Earl et al., 2011; Salzberg et al., 2012), assembly reconciliation was performed using the SGA assembly as 'master' assembly and the SOAP assembly as 'slave' using GARM v0.7 (Soto-Jimenez et al., 2014). Assembly reconciliation is known to improve upon individual assemblies by leveraging their different strengths (Soto-Jimenez et al., 2014; Yao et al., 2012). Gapfilling and scaffolding were run again on the output to refine unresolved regions. The reconciled assembly had increased contiguity over the starting SGA or SOAP assemblies, as measured by a higher N50 statistic (118kb vs. 66kb or 31kb, respectively), a lower number of scaffolds (46,729 vs. 565,630 or 203,468, respectively). The reconciled assembly also had a higher number of complete core eukaryotic genes according to CEGMA (Parra et al., 2007) (223 vs. 219 or 217, respectively; see below). The final assembly has been deposited in Genbank under accession number JNBZ00000000 (first version, referred to as NotFur1 assembly).

Assessment of completeness of the draft genome using CEGMA

The CEGMA algorithm (Parra et al., 2007) was applied to the turquoise killifish draft genome to gain an estimate of the completeness of the genome as well as initial gene models for these core eukaryotic genes (CEGs) (*i.e.* proportion of a conserved eukaryotic core genes present in the draft genome).

De novo transcriptome assembly using Oases

Strand-specific RNAseq libraries from GRZ adult individuals were constructed from liver, brain, testes and tail tissues using the standard Illumina protocol. The liver and testes libraries were polyA-selected, and the brain and tail libraries were rRNA-depleted (Harel et al., 2015). Libraries were sequenced on Illumina HiSeq2000 machines, as paired-end 101bp reads. After read quality preprocessing (see above), libraries were run through FLASH to provide extended merged reads as well as unmerged remaining paired reads. De novo transcriptome assembly was performed using the Oases de Bruijn graph-based algorithm (oases v0.2.06, and velvet v1.2.03 dependency) (Schulz et al., 2012). Assemblies were generated using a k-mer range of 43 to 91 and a step of 4, and keeping only assembled sequences longer than 200bp with supporting evidence of ≥ 5X coverage. Transcriptomes from the four different tissues were assembled separately. Resulting tissue-specific assemblies were then merged, and redundantly assembled transcripts were eliminated using uclust (Edgar, 2010) and cdhit-est cluster (Li and Godzik, 2006), asking for ≥90% reciprocal sequence homology. Resulting transcripts constituted our reference transcriptome assembly.

Higher-order scaffolding using RNA-seq data and RAD-seq linkage map

The raw reads from our Illumina paired-end RNA-seq libraries were used for higher-order scaffolding over potentially unresolved introns. The tophat-fusion pipeline (Kim and Salzberg, 2011) was used to identify read pairs mapping over 2 different scaffolds. Mapping was run for each library as: 'tophat2 --bowtie1 --fusion-search -m 1 -g 5 --fusion-multipairs 1 -o ./OUTPUT-DIR/ --solexa-quals -p 4 -r 100 --mate-std-dev 50 --no-coverage-search --no-mixed --segment-length 55 index_name RNAseq_file_1.fq RNAseq_file_2.fq'. The 'fusion.out' output file was parsed to filter pairs to retain only putative high-confidence links between scaffolds (more than 20 supporting pairs and no contradicting pairs, or more than 50 supporting pairs and less than 1% contradicting pairs). Scaffolding was performed using high-confidence putative links between scaffolds and by stringently filtering out links that could be ambiguous or hard to resolve (e.g., links potentially resulting from alternative splicing). The orientation determined by Tophat for the fusion according to the RNA-seq mapping (forward-forward, reverse-reverse, forward-reverse, reverse-forward) was used to orient the scaffolds with respect to one another. An AGP format file (https://www.ncbi.nlm.nih.gov/assembly/agp/AGP_Specification/) was generated to summarize the scaffold links and placement, and is available for download on the genome browser website.

To perform even higher-order scaffolding on the RNA-seq scaffolded assembly, the linkage map from cross GxM (see below) was used. First, RAD-seq markers were mapped in fasta format to the genome using bowtie1. The alignment file was parsed to extract information on captured genomic scaffolds and on their relationships with mapped markers to build a scaffolding roadmap. During parsing, markers that led to contradictory or ambiguous links that could not be resolved were discarded. A limit resolution confidence threshold of 5cM was used to resolve conflicts arising between assembled genomic contigs and mapping sequence: below 5cM, the genomic order was preferred; above 5cM, the genetic map was preferred. Captured scaffolds were then ordered according to the genetic linkage map using an AGP format output. The rest of the scaffolds were considered 'unplaced' in the assembly and annotated as such in the final AGP output file. This file is available for download on the genome browser website.

Estimate of unresolved repeats in draft assembly

Genomes with a high-repeat content are known to yield relatively fragmented drafts when assembled using short-read sequencing technology (Treangen and Salzberg, 2012). A previous assessment estimated that the turquoise killifish genome contained at least 45% of repetitive sequences (Reichwald et al., 2009). We examined regions of the genome with an excess fold coverage compared to the expected coverage from the libraries (Figure S1A). This was done based on the conservative genome size estimate of 2Gb. A small portion of the assembly captured a large portion of the reads. For example, regions that have 10 times the expected coverage (i.e. mean coverage based on depth of sequencing and used genome size estimate) represent 1% of the genome assembly but capture 17.4% of reads (Figure S1A). Regions with excess coverage likely correspond to unresolved repeats that are present in many copies in the actual genome sequence (Treangen and Salzberg, 2012).

Analysis of repetitive sequences in draft assembly

The annotation of repetitive elements present in the assembly was performed using RepeatMasker v3.3.0 (Smit et al., 1996-2004) and the RepBase repeat library (2012-04-18 version) (Jurka et al., 2005). We used RMBlast version-2.2.27 as an alignment engine, and restricted the similarity search to the library of elements from teleost fishes.

Mapping of Sanger shotgun sequence contigs to the draft assembly

Previously published Sanger shotgun contigs from GRZ genomic DNA were obtained from GenBank (accession ABLO01) (Reichwald et al., 2009). Contigs were aligned to the draft assembly using BLAT (Kent, 2002). The longest alignment of the contig to the genome (the best match) was then used to compute the aligned fraction of each Sanger contig.

Annotation of protein-coding genes using the MAKER2 pipeline

The MAKER2 pipeline was used to generate consensus gene predictions derived from ab initio predicted models, RNA-seq reads, de novo transcriptome assembly and EST/transcript data, and protein similarity (Holt and Yandell, 2011). MAKER2 predicts the most likely gene model and outputs a confidence score (Annotation Edit Distance, AED) to each prediction based on the degree of support by experimental evidence (EST/transcript mapping, RNA-seq mapping, protein homology, etc.). Several sources of transcriptomic sequences were generated to support gene predictions. These include: 1) published assembled transcript sequences from the turquoise killifish downloaded from Genbank (Petzold et al., 2013) and ESTs from another killifish, *H. fondulus* (Tingaud-Sequeira et al., 2013), 2) our de novo assembled transcriptome sequences, and 3) our paired-end Illumina RNA-seq data from brain, liver, testes and tails as well as previously published RNA-seq librairies from skin and whole fish (Petzold et al., 2013). RNA-seq data were aligned to the masked reference genome and exon-exon junctions were modeled using the tuxedo suite (Trapnell et al., 2012), as recommended in the MAKER2 manual. The full-length transcriptome data and splice junctions were used as transcript evidence in the MAKER pipeline. In addition to the transcript sequences, the complete reference proteomes of *Danio rerio*, *Oryzias latipes* and

Takifigu rubripes were downloaded from Uniprot (on 05-30-2013) to provide the MAKER pipeline with protein homology evidence. The MAKER pipeline also incorporates a repeat masking step before running gene predictors, and the *teleostei* repeat library from the RepBase database (2012-04-18 version) (Jurka et al., 2005) was used for this step (see above). The discovery of single-exon genes was enabled, though it is disabled by default in the pipeline, to allow for the discovery of potentially important mono-exonic genes. MAKER aligns transcript and protein evidence to the genome using the sensitive/specific splice site-aware alignment algorithm exonerate (v2.2.0) (Slater and Birney, 2005). Alignments retained by the MAKER pipeline have > 20 score, do not overlap low-complexity regions, and were used to estimate the proportion of transcripts mapping to the assembly with high quality.

The ab initio predictor SNAP (Semi-HMM-based Nucleic Acid Parser) was first trained specifically for the turquoise killifish using CEG models from the output of CEGMA (Parra et al., 2007). After a first run of the MAKER pipeline, gene models with an AED score of 0 (most supported by RNA or protein evidence) were then used to retrain SNAP for a second round to obtain a higher quality hidden Markov model. Ab initio predictor Augustus was then trained using AED=0 gene models from this second run, using the included zebrafish model parameters as a starting point to create a turquoise killifish-specific gene prediction model. We then performed a third and last run of the MAKER pipeline with all evidence support, enabling gene prediction from transcript sources (i.e. exonerate alignment of transcripts and Cufflinks junctions) and from trained ab initio gene predictors (i.e. Augustus and SNAP). In this final run, 61,418 putative protein-coding gene models were predicted by MAKER at any AED threshold (0-1). Gene model filtering steps and annotation of protein coding gene models by sequence orthology are described below.

Analysis of alignment of turquoise killifish transcripts and RNA-seq to the draft genome

Long assembled transcripts from a published catalog (Petzold et al., 2013) and from our de novo assembled transcriptome were used to assess the quality of our genome assembly. Transcript alignments were performed and filtered based on quality using MAKER, which uses exonerate (v2.2.0), an alignment algorithm that is sensitive, specific, and splice site-aware (Slater and Birney, 2005). Alignments retained by MAKER have > 20 score and do not overlap low-complexity regions. They were used to assess the proportion of transcripts mapping with high-quality to the genome assembly. The remaining transcripts that were not part of this high-quality set were then mapped to the genome using BLAT (using the -trimHardA -trimT -extendThroughN options), and transcripts with at least one partial hit to the genome were quantified.

Illumina paired-end RNA-seq reads from adult turquoise killifish tissues (SRP041421) were aligned to our reference draft genome using bowtie 0.12.7 and TopHat2 v2.0.4 (Trapnell et al., 2012). Alignments were retained only if they mapped to at most three loci in the assembly (-g 3). Statistics of percentage of mapped reads and proper pair orientations were obtained from the resulting alignment files using samtools (v 0.1.17).

Alignment of transcripts to proper paralogs for specific genes

For selected genes, we determined independently proper mapping of transcripts to different paralogs (Table S4I). In these cases, the corresponding annotated transcripts from the published catalog of assembled turquoise killifish transcripts (Petzold et al., 2013) were downloaded from the NFIN website (http://nfintb.fli-leibniz.de/nfintb/). Sequences were mapped using BLAT (-trimHardA -extendThroughN -fine options), and alignments were visualized using our genome browser (Table S4I).

Annotation of ribosomal RNA genes and short non-coding RNA genes

RNAmmer (Lagesen et al., 2007) was used to annotate high-quality rRNA genes. tRNAscan-SE (Lowe and Eddy, 1997) was used to annotate high-quality tRNA genes. Putative tRNA with a non-conventional anticodon, or labeled as likely pseudogenes by the software, were discarded from the high-confidence tRNA gene predictions. Annotated tRNA gene type and distribution per anticodon are reported in Table S1A. Infernal (Nawrocki and Eddy, 2013) was used to annotate putative snRNA, snoRNA and miRNA

genes. All programs were run with default settings. RepeatMasker also provided secondary predictions of rRNA, tRNA and snRNA genes, and Infernal provided secondary predictions of tRNA genes.

Annotation of long non-coding RNA genes

Illumina paired-end RNA-seq reads from adult turquoise killifish tissues (SRP041421) were aligned to the reference draft genome using bowtie 0.12.7 and TopHat2 v2.0.4 (Trapnell et al., 2012). Alignments were retained only if they mapped to at most three loci in the assembly (-g 3). De novo transcriptome assembly guided by the genome was performed using cufflinks2 v2.1.1, using the predicted protein-coding gene as prior information, and with a maximum pre-mRNA fraction parameter set at 0.1 (-j 0.1). Previously identified transcripts from predicted protein-coding genes were excluded from subsequent steps.

Next, the EMBOSS software suite was used to predict the longest ORF in each of the remaining predicted transcripts (Mullan and Bleasby, 2002). Transcripts longer than 150bp, with at least 5X sequencing coverage in one library, and whose longest predicted ORF was strictly shorter than 50 amino-acids, were retained for further processing. Finally, as a stringent cutoff, we filtered the putative lncRNA genes to retain only those with an H3K4me3 peak, indicative of promoter activity, in an H3K4me3 ChIP-seq dataset from turquoise killifish adult brain (SRP045718) (Harel et al., 2015).

Ortholog identification and gene model annotation

To further annotate protein-coding genes from all the 61,418 MAKER gene predictions and exclude spurious predictions and untranslated sequences, we used homology-based evidence from 19 fully sequenced genomes (Table S2B), including seven additional fish genomes. Protein sequences for all the organisms were downloaded from Ensembl (release 75) (Cunningham et al., 2014) using BioMart (Kinsella et al., 2011). The sea urchin proteome was downloaded from Ensembl Metazoa. For genes with multiple protein products due to alternative splicing, only the longest protein was used.

An all-against-all BLASTp search was run using an e-value of 10^{-5}. Both best and bidirectional best hits for every turquoise killifish protein in each of the analyzed genome were determined. In addition, a BLASTp search against NCBI nr (all non-redundant protein sequences) was performed using all the predicted turquoise killifish proteins. We discarded 32,924 predicted turquoise killifish genes whose protein product did not have either a match in at least two organisms or a match in NCBI nr with at least 30% query coverage, leading to a final set of 28,494 killifish protein coding genes. These genes were divided into three tiers of confidence levels based on the homology evidence from all of the other genomes. Turquoise killifish genes with bidirectional best hits in at least 10 analyzed organisms were considered Tier-1 (10,329 genes). Turquoise killifish genes with bidirectional best hits in less than 10 organisms, but with homologous sequences in at least 10 genomes, were considered Tier-2 (12,192 genes). Turquoise killifish genes with a BLASTp hit in less than 10 organisms, and a best hit with NCBI nr database, were considered Tier-3 (5,973 genes). Tier-1 and Tier-2 genes represent the high quality gene prediction in the current assembly, and are considered high-confidence protein coding genes (Figure 2B).

Turquoise killifish genes from all the three tiers were assigned a gene symbol based on the consensus symbol from all the genomes having a homologous sequence. In cases where a consensus gene symbol could not be reached, preference was given to the gene symbols supported only by the 7 teleost fish genomes. If still no consensus gene symbol could be identified, the gene symbol was chosen from organisms in the following order of priority: human, mouse, zebrafish, and medaka. For multiple genes with the same gene symbol (possible gene duplicates), a number was assigned randomly to each ('NofN'). Finally, an ID was assigned to each protein (Figure S1G) and long non-coding RNA genes (Figure S2E). This ID contains information about the scaffold, synteny, and number of organisms having a homolog of the gene, or for the lncRNA, the tissues where the gene is expressed.

Gene family analysis

To construct gene families in turquoise killifish and medaka, an all-against-all BLASTp was run using all the filtered turquoise killifish and medaka proteins (e-value < 10^{-5}). The hits were then clustered at 5

similarity thresholds (50 to 90) to generate gene families using TransClust (Wittkop et al., 2010). Using the resulting families with at least one gene from both the organisms, we performed linear regression without an intercept term in R. As a control, an identical analysis was performed between platyfish and medaka proteins. A good correlation genome-wide between gene families indicates proper assembly of paralogs. Differences in the paralog numbers of specific families between killifish and medaka could be due to evolutionary events such as lineage specific gene duplication or to loss misassembly/misannotation in one of the species.

Computation of codon usage

All high-quality annotated protein-coding genes (Tiers 1, 2 and 3) were used to estimate codon-usage in the turquoise killifish. The R (http://cran.r-project.org) package 'seqinr' (Charif and Lobry, 2007) was used to compute codon usage from the corresponding coding sequences (Table S1B).

Metagene profiles for RNA-seq and metapromoter profile for H3K4me3 ChIP-seq

For metagene analysis of RNA-seq, normalized aligned read counts were extracted around the annotated TSSs (-tss option), gene bodies (-rna), or TTSs (-tts) using the 'annotatePeaks.pl' script from the HOMER suite (Heinz et al., 2010) with '-hist' option and with the final annotation gff3 file for the turquoise killifish genome. Average values are used to plot the metagene profiles.

For metapromoter analysis of H3K4me3 ChIP-seq, normalized aligned read counts were extracted around the annotated TSSs (-tss option) using the same method as above. Average values are used to plot the metapromoter profile.

Analysis of RNA-seq data for transposase expression

Our de novo assembled transcriptome was used to obtain transcript reference sequences derived from independent transposon insertions of the same family. We mapped reads to the transcriptome instead of the genome because genome-based quantification of RNA expressed from transposable elements is problematic due to ambiguous mapping of reads to the genome. Transposon-derived transcripts were annotated using BLASTp hit to the NCBI nr database or using predicted domains from the NCBI conserved domains database (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml). The RNA-seq reads were mapped to the reference transcriptome using Tophat2, supported by bowtie1 (Trapnell et al., 2012). Aligned reads over transcripts were counted using the bedtools suite 'coverageBed' command. Read counts were normalized by transcript length and library size to obtain final FPKM values.

Synteny analysis

To build whole genome synteny maps, genomic scaffolds were assigned to specific RAD-seq markers along the linkage map using BLAT. Linkage groups for the turquoise killifish were then represented as a linear series of scaffolds matching the RAD-seq markers (not oriented). Coding sequences matching each mapped scaffold were then extracted using a GTF file of the turquoise killifish annotated genes, and a new file was built containing gene names along the mapped scaffolds. Medaka, stickleback, and platyfish GTF files were obtained from ftp://ftp.ensembl.org/pub/. The ordered linkage groups in the turquoise killifish were matched to the medaka and platyfish chromosomes based on reciprocal best BLASTp hits (turquoise killifish/medaka synteny), or on identical gene names (platyfish/medaka synteny). High quality protein genes from Tier 1 and Tier 2 were used for the turquoise killifish. Matching pairs of genes between species were plotted using either the turquoise killifish linkage map position or the platyfish chromosome position on the y-axis and the medaka chromosome position on the x-axis (OxGrid plot).

Construction of the species tree

For the generation of the species tree, we selected all the 619 one-to-one orthologs from each of the organisms where the same killifish gene was the reciprocal best BLASTp hit in each of the 19 animal genomes. These genes represent the entire set of high confidence one-to-one orthologs for the 20 organisms, including turquoise killifish, in our analysis (Table S2A). Known aging-related genes are well represented in this list (see Table S2C). A single long sequence for each organism was constructed by concatenating these 619 proteins in the same order. The sequences were then aligned using MAFFT (Katoh and Standley, 2013). The conserved blocks were identified from the corresponding multiple sequence alignment to remove non-conserved or misaligned regions using Gblocks (Talavera and Castresana, 2007). ProtTest (Darriba et al., 2011) was used to identify the best model (LG+G+I) for construction of phylogeny. A maximum likelihood tree was then constructed using PhyML v3.1 with 100 bootstrap steps for statistical support (Guindon et al., 2010). We used discrete Gamma model with 4 categories (shape parameter: 0.821) and allowed PhyML to estimate the proportion of invariant sites (0.066). The resulting unrooted tree was rooted in MEGA-6 based on *C. elegans* as the outgroup. Phylogeny based on neighbor-joining (Tamura et al., 2013) produced identical topology.

Phylogenetic trees for specific gene families

The predicted protein sequence corresponding to a specific turquoise killifish gene was used to identify best BLASTp hits in the proteomes of 7 fish species. Hits for medaka, stickleback, *Tetraodon, Takifugu,* cod and zebrafish were obtained from the UCSC/ENSEMBL portals if available; if not, the best BLASTp hits from NCBI nr database were used. The best hits for platyfish protein sequences were always obtained using BLASTp hits against NCBI nr.

Protein sequence alignment was performed using ClustalX v2.1. PHYLIP proml v3.695 was used to build trees, using zebrafish sequences as the outgroup. The plots were generated as rooted phylograms using Unipro UGENE v1.17.0.

Identification of turquoise killifish genes under positive selection

To identify genes in the African turquoise killifish lineage that are under positive selection, we used PAML (Phylogenetic Analysis by Maximum Likelihood, version 4.8) (Yang, 2007; Yang and Nielsen, 2002), which implements a maximum likelihood framework to evaluate adaptive selection based on non-synonymous by synonymous substitution rate ratio (K_A/K_S, D_N/D_S or ω ratio). We used a branch-site model implemented in PAML to identify individual amino-acid sites targeted by positive selection in the turquoise killifish branch using 7 other long-lived fish species as a background (Figure S3A).

First, single ortholog gene families were selected from the eight teleost fish genomes (including the turquoise killifish) where a clear bidirectional best hit with the same turquoise killifish gene was present in at least four other fish genomes. Hence, we required at least 5 fish sequences including the turquoise killifish for further analysis, leading to a set of 13,637 ortholog families. These sequences were aligned using PRANK (Loytynoja and Goldman, 2005) (codon model incorporated in GUIDANCE (Penn et al., 2010)), an algorithm known to generate highly accurate alignments to detect positive selection (Fletcher and Yang, 2010; Jordan and Goldman, 2012). Since alignment quality is one of the critical steps in accurate detection of positive selection, we applied stringent filtering using GUIDANCE (Penn et al., 2010), which is known to be a highly accurate alignment quality-control package for such analysis (Jordan and Goldman, 2012). The following (local and global) stringent filters were applied on each alignment using GUIDANCE: mean residue pair score > 0.85 (indicative of the overall alignment quality), mean column score > 0.85 (indicative of the overall alignment quality), no individual sequences with score < 0.85 (indicative of the sequences that may lead to bad alignment quality) and no sequence-pairs with score < 0.85 (indicative of the sequence pairs leading to bad alignment).

For the alignments that passed our quality filters, we used the branch-site model implemented in CODEML that is designed to detect positive selection that affects only a few sites on the specified branches of a phylogeny. On the species tree of the eight fish genomes, the turquoise killifish lineage was marked as 'foreground' and the rest of the fish species as 'background' lineages (Yang, 2007; Yang and Nielsen,

2002). We then performed a likelihood ratio test between model M2a_null (model = 2, NSsites = 2; fix_omega = 0) and M2a_selection (model = 2, NSsites = 2; fix_omega = 1, omega = 1) as recommended (PAML User Guide: http://abacus.gene.ucl.ac.uk/software/pamlDOC.pdf).

A p-value to assess the significance of the likelihood ratio test was calculated using χ^2 test for twice the difference of likelihood from model M2a_selection versus likelihood M2a_null, with one degree of freedom. We used Bayes Empirical Bayes (BEB) probabilities provided by CODEML to identify individual sites under selection (Yang et al., 2005). However, we further excluded the selected sites if the GUIDANCE column score for the site under selection was less than 0.85 or if there was a gap in the alignment (in any sequence) within +/- 5 amino-acids from the selected site, even if the BEB probability value was significant. This stringent filter was introduced to ensure that the sites under selection are in the well-aligned regions. Finally, we subjected the p-value of the likelihood ratio test to multiple hypothesis correction in R by Benjamini–Hochberg method and identified the genes with significant overall p-value at 5% FDR ($p < 0.05$). There were 497 genes under positive selection with a corrected overall p-value < 0.05 (Table S3B). To generate an even more stringent list, we restricted the genes to those with at least one site with BEB probability of selection > 0.95, which led to 249 genes (Table S3A). The 'highest confidence' list of 249 genes was used for GO enrichment analysis, functional effect prediction and overlap with genes under selection in extremely long-lived vertebrates. The list of 497 genes was used to assess the overlap with aging genes in vertebrate model organisms from the GenAge and LongevityMap databases.

While we designed our analysis to identify the genes and residues under positive selection with high confidence, there may be additional genes/sites under selection that are missed or some false positives due to inherent limitations of this kind of analysis (e.g. misalignment of codons, choice of background species, missing exons due to alternative splicing, parameter sensitivity of likelihood ratio calculations, etc.). In general, the false positive rate of branch-site tests ranges from 4 to 6.4% (Wong et al., 2004; Yang and dos Reis, 2011).

GO enrichment analysis of genes under positive selection

Gene Ontology (GO) terms for the predicted killifish genes were obtained by attributing the corresponding GO terms of zebrafish genes to their one-to-one orthologs in the turquoise killifish genome. GO enrichment analysis for the stringent list of 249 genes under positive selection was performed in R (version 3.1.1) using 'GOstats' package (Falcon and Gentleman, 2007). We used all the filtered 13,637 genes as background and employed hyper-geometric test implemented in GOstats to obtain the significantly enriched terms after Benjamini-Hochberg correction for multiple testing, and filtering out the terms that showed significant depletion from the results. Selected GO terms with p-values, number of genes, and enrichment values ≥ 2 fold from the top enriched GO terms are shown in Figure 3C.

Overlap between genes under positive selection in the turquoise killifish and genes that change in expression with age

For analysis of the publicly available brain aging dataset (MZM-0410 strain) (Baumgart et al., 2014), the STAR ultrafast universal RNA-seq aligner (Dobin et al., 2013) was used. Reads over gene models were counted using the featureCounts feature of Subread package (Liao et al., 2014). Library size adjustment and dispersion normalizations were performed using the R 'DEseq2' package (Love et al., 2014). DESeq2 was used to model expression changes as a function of the age of the fish (5, 12, 20, 27 and 39 weeks). Differentially expressed genes according to that model were called at FDR threshold of 5%. The list of differentially expressed genes was overlapped with the list of 249 positively selected genes in the turquoise killifish. Enrichment was measured by a Fisher test on expressed genes. Expression levels of genes of interest were plotted as a heatmap using the 'pheatmap' package in R (http://cran.r-project.org/web/packages/pheatmap/index.html).

Turquoise killifish orthologs of aging-related genes from GenAge and LongevityMap

Aging-related genes in mouse or human were downloaded from the GenAge database (Built 17) (de Magalhaes et al., 2009; de Magalhaes and Toussaint, 2004). Genes with longevity variants in human were obtained from the LongevityMap database (Budovsky et al., 2013) after excluding all the non-significant variants. We used the turquoise killifish orthologs from either the mouse or human genes in the all-against-all BLASTp analysis from all 20 organisms, as described above. These lists are included in Table S4A.

Prediction of functional impact of turquoise killifish variants

To determine the residues under positive selection in the turquoise killifish or the non-synonymous variants between killifish strains that have a functional impact, we used two sequence-based prediction algorithms: PROVEAN (Protein Variation Effect Analyzer) (Choi et al., 2012) and SIFT (Sorting Intolerant from Tolerant) (Kumar et al., 2009). For the residues under positive selection in the 249 genes that had a medaka ortholog in our analysis, we used the medaka protein sequences as reference, and evaluated the potential impact of their substitution to the turquoise killifish specific residue. We then calculated the SIFT score using SwissProt database and PROVEAN score using the NCBI nr database. A replacement was classified as 'DELETERIOUS' by SIFT if the prediction score is < 0.05. For PROVEAN, a score < -1.3 was considered to have a 'Moderate' effect and a score < -2.5 was considered to have a 'Deleterious' effect. After removing potential false positives, we could obtain a functional prediction for 1509 residues under positive selection in 199 genes from SIFT or PROVEAN (Tables S3D and S4D).

To have a neutral reference during the prediction of functional effect using SIFT and PROVEAN, we also predicted the effect of residues under positive selection using the ancestral sequences for the 7 aging genes [BAX, IGF1R(1of2), INSRA, IRS1(2of2) XRCC5, LMNA(3of3) and MGAT5(1of3)]. The sequences of the common ancestors of medaka, platyfish and killifish were generated using maximum-likelihood approach in PhyloBot (http://PhyloBot.com) using zebrafish/cod/stickleback as outgroups. For LMNA(3of3) and MGAT5(1of3), medaka was not a part of our analysis because there was no bidirectional best hit to a medaka ortholog. Therefore, in these two cases, we used the sequence of the common ancestor of platyfish, killifish, fugu, and tetraodon. We then computed the functional impact of the killifish residues at the corresponding position using ancestral sequence from both SIFT and PROVEAN (Table S4E).

To determine the potential functional impact of non-synonymous variants in aging-related genes between killifish strains, we used either the GRZ variant or the variant that is common between MZM-0403 and MZM-0703 ('MZM') as reference points, with the other serving as alternative sequence. This yielded 4 predictions for each considered position: effect of the MZM residue in the GRZ protein sequence context according to SIFT (1) or PROVEAN (2), and effect of the GRZ residue in the MZM protein sequence context according to SIFT (3) or PROVEAN (4). We considered that there was a predicted functional impact if at least one of these predictions was significant. The impact was considered higher confidence if both SIFT and PROVEAN predicted an impact on protein function.

For a subset of aging and age-related genes (those for which we could map the residues under positive selection on the available protein structure, e.g. BAX, IGF1R(1of2), INSRA, XRCC5, and GRN), we used five different methods to assess protein folding/stability changes upon point mutations: ENCoM: http://bcb.med.usherbrooke.ca/encom.php (Frappier et al., 2015), PoPMuSiC 3.1: http://dezyme.com/ (Dehouck et al., 2011), I-Mutant v3.0: http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant3.0/I-Mutant3.0.cgi (Capriotti et al., 2006), DUET: http://bleoberis.bioc.cam.ac.uk/duet/ (Pires et al., 2014), and CUPSAT: http://cupsat.tu-bs.de/ (Parthiban et al., 2006). All these methods model point mutations on protein structure and determine if the free energy of folding ($\Delta\Delta G$) is significantly different between the 'mutant' (in this case, turquoise killifish residue) and 'wild-type' (in this case, human or mouse residue in the structure). If the $\Delta\Delta G$ value was greater than 1 between wild type and mutant, the mutation was considered to have a strong effect on the folding and stability of the protein. If the $\Delta\Delta G$ was > 0.5, it was considered to have a moderate effect (Tables S4F and S7H).

Mapping residues and variants on available protein structures or domains

Residues under positive selection in the turquoise killifish were mapped on the available three-dimensional protein structures for the orthologs of 4 aging genes: INSRA, IGF1R(1of2), XRCC5, and BAX. We first performed a structure-based sequence alignment using the corresponding chain in the available crystal structure in Protein Data Bank (PDB; see accession numbers in Table S4F). To this end, we used all the fish species that were part of our selection pipeline, including the ancestral sequences generated by PhyloBot (http://PhyloBot.com), and aligned them with the corresponding chain in PDB structure using PROMALS3D (Pei et al., 2008). To map the non-synonymous variant with functional effect in GRN (W449), we aligned only the corresponding GRN domain in the turquoise killifish with the PDB structure of the human ortholog (Table S7H) using PROMALS3D. Corresponding residues were then mapped and highlighted on the PDB structures using JalView (Waterhouse et al., 2009) and PyMOL: https://www.pymol.org/ (see Tables S4F and S7H for details).

There were no protein structure available for the orthologs of MGAT5(1of3) and IRS1(2of2), and the available structure for LMNA did not encompass the orthologous residues under selection. Therefore, for these proteins, we mapped the residues under selection on the predicted domains from the NCBI Conserved Domain search (Marchler-Bauer et al., 2015) (CDD: http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi).

To map human variants, we obtained the position of residues associated with exceptional human longevity from LongevityMap (Budovsky et al., 2013; Suh et al., 2008) (IGF1R), and (Conneely et al., 2012; Sebastiani et al., 2012) (LMNA). We aligned the human sequence with multiple fish sequences using PRANK and mapped the residues on the turquoise killifish ortholog. We obtained the residues mutated in LMNA in Hutchinson Gilford Progeria Syndrome via OMIM: http://omim.org/entry/150330 and UniProt: http://www.uniprot.org/uniprot/P02545. We also mapped residues with unique amino-acid changes in 34 genes in the bowhead whale (Keane et al., 2015) (Table S4G) and residues with unique amino-acid changes in IGF1R in the Brandt's bat (Seim et al., 2013). *daf-2* alleles with the phenotype 'extended life span' in *C. elegans* were obtained from WormBase (http://wormbase.org, Release:WS249; Date: September 3, 2015). Amino-acid changes in DAF-2 were from (Patel et al., 2008) and were mapped to IGF1R(1of2) in the turquoise killifish.

We obtained GRN mutations associated to neurodegenerative diseases in humans from OMIM (http://www.omim.org/entry/138945) and from the 'Alzheimer Disease & Frontotemporal Dementia Mutation Database' (Cruts et al., 2012). After manual inspection of the multiple protein alignments of human and fish sequences, we mapped the diseases mutations to the NMR structure of GRN using alignment by MUSCLE. Multiple alignments for ZNF800A and IFI35 were also performed using MUSCLE.

Targeted Sanger re-sequencing of selected residues from aging-related genes

Genomic DNA from fish tissues was extracted using 200 µL of DirectPCR Tail (Viagen Biotech Inc) with 4 µL of Proteinase K (Invitrogen Inc). Samples were incubated at 50°C overnight, boiled at 100°C for 10 min, and centrifuged at 8,000g for 5 min. The supernatant was directly used as a template for PCR reactions.

To confirm the sequence of genes of interest around the residues that were identified to be under positive selection according to the reference GRZ genome, small amplicons (300-400bp) encompassing the residues of interest were amplified by PCR, using the GoTaq Green Master Mix (Promega) with an annealing temperature of 58°C. Two independent GRZ individuals were used for each analysis.

The following primers were used for PCR amplification of genomic sequences:

IGF1R_1of2_V121-T126_F	TTTGCCCTAACACATCTCCATTC
IGF1R_1of2_V121-T126_R	GTGATGTTCCTCAGGTTGTACAG
IGF1R_1of2_F351_F	TTCTGAAACTGACCTCTTCACCT
IGF1R_1of2_F351_R	AGTGTACGTCCTTACCGATTAGT
IGF1R_1of2_A391_F	CGTTCCTCTGTCTACCTCAGTGT

IGF1R_1of2_A391_R	TGAAGTGTACGTCCTTACCGATT
IGF1R_1of2_L426-H428-L429_F	ACCACTGCTAGAATTTTCAGACG
IGF1R_1of2_L426-H428-L429_R	TCTTCTCCCAAACTTTCCAGAGA
IGF1R_1of2_Y489_F	CCATGTTTGTGACCATCTGATG
IGF1R_1of2_Y489_R	CGACACCATGAAATGAGAAAAGC
IGF1R_1of2_A843_F	CCTCGCTATATTTCACTGTTTGG
IGF1R_1of2_A843_R	AAGTTTGCACATTTCCATCAAGT
IGF1R_1of2_L1008_F	AATCTGAGCCCAGGAAACTATTC
IGF1R_1of2_L1008_R	CAGAAGCACTCACCTCTTTTTGT
IGF1R_1of2_L1305_F	GTTGTATTGTGGGAAATTGCCAC
IGF1R_1of2_L1305_R	CTCTATTGGCCAACACTGATGAG
IGF1R_1of2_S1374_F	CCCATTTCAGGGAAGTAAGTTTC
IGF1R_1of2_S1374_R	ATTCATGTGTGCGTATGGTTGTA
INSRA_N425_F	AACACAAGCTAAGCATCCTGAAG
INSRA_N425_R	CAGCTTCATCACGTCTATGTTCA
INSRA_S434-P435_F	TGAAAGCGTGCTGATTCAAATTG
INSRA_S434-P435_R	TGCAACAAAGTCTAGCGATTTCT
INSRA_V457-V459_F	TTTCCCCTCCAGTTTATTTTCAT
INSRA_V457-V459_R	GCAAAGCAACAGTTTGGTCTAGT
INSRA_N566_F	CTGATGGAGACAAAGAGCGTATT
INSRA_N566_R	TCCGTGATTGACTTCCTGTTTAT
INSRA_A705_F	TAAATTCTCCCAAGTTCTGGACA
INSRA_A705_R	GGGTTTTCGCTTTTATTCAAGAT
INSRA_P801_F	AACCCCAGATGTTTTTGGTTACT
INSRA_P801_R	TATAGGATGATGATGCCGTTAGG
INSRA_V910_F	TAATTGTACCTGAATGGGGTGAC
INSRA_V910_R	TACTCAGGGTTTGAAGAGGCATA
INSRA_T1258_F	TTGTTCTGGAGCTCTAACTCACC
INSRA_T1258_R	CAGTATTCGTCCATTGGTCTTTC
XRCC5_G888_F	ACCCCTTTAACTTTGTTCTCAGC
XRCC5_G888_R	ATACAGAAAGTGGTCCGTTTCAC
IRS1_2of2_L94_F	ATGATTCAGTTGTAGCCCAGAAC
IRS1_2of2_L94_R	GCAATCCCATCCTCACCTTTC
LMNA3_M307_I358_F	ATTTGAGAGCAAACTGGCAGA
LMNA3_M307_I358_R	AGTAAAAGCTGTCCGAGTACCTG

The presence of single PCR products was assessed on 1% TAE agarose gels, and PCR products were purified using the Qiagen PCR purification kit. Sequence of individuals was determined by direct Sanger sequencing of PCR products using the original amplification primers (MCLAB sequencing services). When direct sequencing of the PCR products did not work readily, the PCR products were cloned into the PCR4 TOPO-TA vector (Lifetechnologies), and at least two independent bacterial clones from each PCR were sequenced using the M13 reverse primer. Sequence analysis was done using pairwise BLAST alignments to the reference genomic sequence. All the tested PCR amplicons had sequences identical to the genomic reference around the residues of interest.

Genetic variation in GRZ, MZM-0703, and MZM-0403 individuals

To identify genetic variants among different strains of the turquoise killifish, we used next-generation sequencing to genotype the founders of cross GxM (GRZ female and MZM-0703 male). Ethanol-preserved

tissues were used for genomic DNA extraction, library construction (200bp inserts), and paired-end sequencing on Illumina HiSeq2000 instruments (BGI).

Library name	Strain	Insert size	Strategy	Total QC length (bp)	Coverage (X)[a]	Sex
P0GFA	GRZ	200	Paired-end	38,540,978,200	19.27	F
P0GMA	MZM-0703	200	Paired-end	107,816,807,600	53.91	M

[a] coverage estimate based on a conservative genome size estimate of 2Gb.

A library using genomic DNA from a male from another wild-derived strain, MZM-0403, was constructed at Stanford University and sequenced at the Stanford Genome Center.

Library name	Strain	Insert size	Strategy	Total QC length (bp)	Coverage (X)[a]	Sex
MZM0403	MZM-0403	300	Paired-end	12,817,270,468	6.41	M

[a] coverage estimate based on a conservative genome size estimate of 2Gb .

Quality filtered and trimmed paired-end sequencing reads were aligned to the reference genome using bwa v0.6.1-r104, which is a sensitive aligner recommended for variant calling from next-generation sequencing data (Liu et al., 2013). The GATK genotyping pipeline (McKenna et al., 2010), which was developed for human variant calling in the 1000 genome project, was used to call SNPs. GATK v1.6-13 was used with underlying support from picard-tools v1.55. Briefly, PCR-duplicates were filtered out, indels were fine-realigned, base phred scores were recalibrated, and the unified genotyper was run. To recalibrate SNP scores, we assembled a benchmark SNP database using turquoise killifish SNPs from dbSNP (Kirschner et al., 2012) mapped to our genome as well as high quality and a catalog of high-depth SNPs identified by RAD-seq. GATK filters used as parameters to "-A" were: AlleleBalance, DepthOfCoverage, HomopolymerRun, QualByDepth, and MappingQualityRankSumTest. To minimize the rate of false negative calls in the library from the GRZ individual (i.e. sites present but detected with lower depth), we did not apply a sequencing depth cutoff in the GRZ individual (i.e. DepthOfCoverage filter), but we applied sequencing depth cutoffs for the MZM-0703 individual (Depth ≥ 20) and MZM-0403 individual (Depth ≥ 5, to take into account the overall lower sequencing depth of that sample). Only variants with mapping quality ≥ 4 and depth ≥ 5 and low strand bias were selected, which eliminates variants that are hard to validate. Variants with non-straightforward allelic frequencies (i.e. not clearly homozygous or heterozygous) were also eliminated, as these might be artifacts of poorly resolved repetitive regions. Variants annotated as 'LowQual' by GATK were also eliminated. Variants called independently in more than one library are considered as higher quality variants.

To annotate the identified variants and predict their potential function, we used the SNPeff pipeline (Cingolani et al., 2012) on final vcf call files and the high quality predicted gene models (Tiers 1, 2 and 3) gff3 file. The SNPeff pipeline categorizes genetic variants based on their location with respect to genes (e.g. upstream (-5kb from the annotated TSS), downstream (+5kb from the annotated TTS), intronic, etc.), and their potential impact on the coding sequence (e.g. synonymous, non-synonymous, etc.).

Genetic crosses

One female from the GRZ strain was crossed with one male from the MZM-0703 strain (cross GxM), and one female from the Soveia strain was crossed with a male from the GRZ strain (cross SxG). F1 fish were interbred in families to generate F2 fish.

	F1	F1 families	F2
Cross GxM	36	16	430
Cross SxG	9	4	130

Observed lifespan was scored as the age at death for all the fish. For this study, fish were raised in cohorts of mixed sexes, in the conditions described above. Note that both genetic and environmental factors affect the age of death. Casualties due to non-natural death causes, e.g. inter-individual aggression or occasional tank exclusion from the water recirculation system, were not scored as "observed lifespan" and these individuals were censored for the lifespan analysis. Survival analysis was done with the R package "survival" using Logrank test statistics.

RAD-seq library construction for Illumina sequencing

RAD-seq libraries for 225 samples from the cross GxM and for 86 from the SxG cross were prepared as described (Etter and Johnson, 2012). For the libraries, either 125 or 200 ng of genomic DNA from each sample was digested for 60 min at 37^{o}C in a 25 µL reaction volume containing 2.5 µL 10x Buffer 4 and 10 units (U) SbfI-HF (New England Biolabs [NEB]) in a 96-well PCR plate. Samples were heat-inactivated for 20 min at 65^{o}C then allowed to cool at room temperature for 1 hour. 1.0 µL or 1.6 µL of 6bp barcoded SbfI-P1 Adapter (100 nM), a modified Illumina © adapter (2006 Illumina, Inc; top oligo: 5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTxxxxxxTGC*A-3'; bottom oligo: 5'-Phos-yyyyyyAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT-3' [x and y denote barcode and reverse complement, respectively]), was added to each sample along with 1.4 µL rATP (25 mM, Epicentre), 0.4 µL 10x NEB Buffer 4, 0.25 µL (500 U) T4 DNA Ligase (high concentration, NEB), 1.95 µL H_2O and incubated at room temperature for 30 min. Samples were heat-inactivated for 20 min at 65^{o}C and cooled at room temperature for 30 min, then combined in sub-libraries of 24-30 F2 individuals (15 µL per F2 library) and processed as 4 parallel libraries. 180 µL of each pooled sample was randomly sheared (Bioruptor) to an average size of 500 bp. 30 µL sheared sample was run on a 1.25% agarose gel to determine size before a 1.0X AMPure XP bead size selection and purification was performed on the remaining sheared volume. The Quick Blunting Kit (NEB) was used to polish the ends of the DNA in a 25 µL reaction volume containing 2.5 µL 10x Blunting Buffer, 2.5 µL dNTP Mix and 1.0 µL Blunt Enzyme Mix incubated at room temperature for 30 min. The sample was purified with AMPure XP beads, including a 0.5x size-exclusion step prior to 1.0X purification, and incubated at 37^{o}C for 20 min with 10 U Klenow Fragment (3'-5' exo-with 2.5 µL NEB Buffer 2 and 0.5 µL dATP (10 mM, Fermentas), to add 3' adenine overhangs to the DNA. The reaction was moved to room temperature for 30 min, purified (1.0X) and 1.0 µL of Paired-End-P2 Adapter (PE-P2; 10 µM), a divergent modified Illumina © adapter (2006 Illumina, Inc,; see (Etter et al., 2011)) was ligated to the DNA fragments. The sample was purified (1.0x), eluted in 50 µL, and quantified using the QubitTM. We used 2.1-4.2 ng equivalent per individual as template in a 50 µL PCR amplification with 25 µL Phusion Hot Start Flex 2X Master Mix and 2.0 µL modified Illumina © amplification primer mix (10 µM, 2006 Illumina, Inc.; long-P1-forward primer: 5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T-3', short-P2-reverse primer: 5'-CAAGCAGAAGACGGCATACG*A-3'). PCR was carried out with an initial denaturing step at 98^{o}C for 3 min, then 14 cycles of 40 sec at 98^{o}C, 15 sec at 65^{o}C, and 30 sec at 72^{o}C followed after 14 cycles by 5 min at 72^{o}C. RAD-PE GP libraries were prepared together from 225 ng of genomic DNA and 2.0 µL of P1 adapter each and processed as above as a sub-library except that a gel extraction was performed after PCR and bead cleanup to enrich for longer fragments as previously described (Etter et al., 2011). Libraries were purified, diluted to 10 nM and submitted to the University of Oregon Genomics Core Facility for qPCR quantification. Sub-libraries were mixed at equal molar quantities and sequenced on the HiSeq 2000 following Illumina protocols for 100bp single-end reads (paired-end in the case of the GP library). All sequences are available at the NCBI Short Read Archive (accession number: SRP041421).

RAD-seq analysis of genetic cross

RAD-seq data from the GxM genetic cross was processed and analyzed using STACKS (Catchen et al., 2011). We identified a total of 65,773 RAD-tags. Tags containing only one SNP were selected (-F snp_l=1; -F snp_u=1), corresponding to 9,529 in cross GxM. RAD-tags represented in less than 25% of the genotyped subjects were excluded from the analysis, resulting in 8,399 markers.

Linkage map generation based on the RAD-seq data

We built a linkage map with R/qtl using RAD-seq markers that had homozygous haplotypes in the grandparents. This map was used for genome scaffolding and identification of QTL. Individuals that had genotypes in <1400 markers were dropped. Markers were considered duplicates and removed if >80% of pairs of individuals had matching genotypes. Pairs of individuals with >80% matching genotypes were also removed. Markers with distorted segregation patterns ($p<10^{-10}$) were also dropped. Genotype frequencies were examined and confirmed to be 1:2:1 AA:AB:BB. The final cross GxM linkage map comprised 193 F2 individuals and 5,757 RAD-seq markers.

Random Forest QTL mapping

RAD-seq markers with a minor allele frequency below 25% were removed, i.e. 2,672 markers, corresponding to 32% of the total markers. Additionally, individuals for which genotyping information was not available from more than 25% of the remaining markers were excluded for this analysis. This corresponded to 53 individuals, i.e. 24% of the total number of fishes. To test the effect of sex on longevity, QTLs were mapped in: i) all individuals, ii) males only, iii) females only, iv) the residual of the regression of the longevity data on sex; furthermore genetic association with v) weight, vi) weight in males, vii) weight in female, viii) the residual of the regression of weight on sex, as well as other discrete traits: ix) gender, x) tail color, and xi) presence of a black strip.

QTL detection was carried out using a method based on Random Forest (Michaelson et al., 2010), that we previously adapted to take missing values and population structure into account (Clement-Ziza et al., 2014). This method uses genetic markers as predictors to model the traits and population structure is modelled as covariates. First, we estimated the kinship matrix, which scores the relatedness between the strains. We removed markers with more than ten missing genotype values for the population structure estimation. Missing values were randomly replaced by random alleles with probabilities following the distribution of the alleles for the marker of interest. The procedure was repeated 2,000 times, each time building a new kinship and performing singular value decomposition. The average of all generated matrices was used as the final estimate of the kinship. As covariate, we selected the eigenvectors corresponding to the top five eigenvalues, which accounted for more than 75% of the genotype variance.

For the QTL mapping, forests of 14,400 trees were grown (120 forests of 120 trees) using the R implementation of the 'RandomForest' algorithm. The strategies described previously (Clement-Ziza et al., 2014) and above were used to handle missing genotype values and model the population structure. The *mtry* parameter was left to default (one third of the total number of predictors). The QTLs were then scored using the predictor selection frequency as previously proposed (Michaelson et al., 2010). To estimate the significance of the linkages, each trait was permuted 50,000 times and random forests were trained for each permutation. The correspondence between the covariates and the permutated traits was maintained in order to properly estimate the significance of the trait-marker linkages. We obtained null distributions of the selection frequencies for each trait and each marker. They were used to estimate p-values for the selection frequencies. We then reused the permutation results to also estimate p-values for each randomized trait and thus obtained a null distribution of p-values. We then considered the 11 mapped traits together to estimate the false discovery rates (FDR) for the entire analysis based of null distribution of the p-values.

Identification of genomic scaffolds underlying the lifespan QTL

Bowtie 0.12.7 was used to map the 6 RAD-seq markers corresponding to the lifespan QTL peak on LG-3 to our assembled scaffolds. Scaffolds were considered to belong to the peak if at least one of these RAD-seq markers uniquely mapped to them. This analysis resulted in 6 scaffolds (Table S7). Of note, two of these scaffolds (GapFilledScaffold _60 and GapFilledScaffold_883) also contain RAD-seq markers that were attributed to other linkage groups by R/qtl (marker 42243 was attributed to LG-2 and markers 18875 and 24430 were attributed to LG-15), although the genes in GapFilledScaffold_883 showed synteny with the equivalent region in medaka that is syntenic to LG-3. To be comprehensive, both scaffolds were kept in the analysis. Protein-coding genes and non-coding RNA genes contained in these scaffolds were then identified from our annotation files. Gene positions were plotted to scale ordered on the scaffolds using the 'grid' R

package. SNPs between the GxM cross founders falling on these scaffolds were extracted for further analysis. Circular linkage maps and scaffolds were plotted using Circos: http://circos.ca/ (Krzywinski et al., 2009).

Analysis of enrichments in aging genes at the lifespan QTL

To compare the potential enrichment in aging genes at the lifespan QTL with the rest of the linkage map, we first extracted all the scaffolds that contained RAD-seq markers as described above. Genes belonging to these scaffolds were attributed to the corresponding linkage groups (LGs). There was a total of 13,242 genes anchored to the linkage map.

To test the enrichment of aging-related genes at the lifespan QTL, we used the list of combined human and mouse aging and longevity-related genes from the GenAge database (Table S4A). Out of the 462 genes in the mouse and aging complete list, 238 were anchored to our linkage map, including 20 in LG-3, and 5 at the QTL peak. Under the hypothesis of random distribution of genes, we applied Fisher's exact test to measure enrichment of aging-related genes at the QTL compared to the entire linkage map or just LG-3. We also used the list of mouse aging and longevity-related genes from GenAge (Table S4A). Out of the 142 genes in this mouse list, 70 were anchored to our linkage map, including 6 in LG-3, and 3 at the QTL peak.

2010 fish collection expedition to Mozambique

To identify genetic variants present in wild turquoise killifish populations, we conducted an expedition and collected wild specimens from 5 different localities along the Chefu river drainage in southern Mozambique in April-May 2010.

Strain name	GPS coordinates	Male coloration
ZMZ-1001	S21˚ 48.933' E031˚ 55.872'	Yellow
ZMZ-1002	S21˚ 55.011' E021˚ 05869'	Yellow
ZMZ-1003	S22˚ 08.803' E032˚ 49.465'	N.A.*
ZMZ-1004	S22˚ 28.924' E032˚ 49.465'	N.A.*
ZMZ-1005	S22˚ 30.497' E032˚ 33.055'	Yellow; Red
ZMZ-1006	S23˚ 27.548' E032˚ 33.855'	Red
ZMZ-1007	S24˚ 06.293' E032˚ 46.117'	Red

*only females were identified, therefore male coloration could not be assessed

Targeted Sanger re-sequencing of candidate genes in individuals from different strains or from the wild

Genomic DNA was extracted as described above. To genotype the *GRN*, *IFI35* and *ZNF800A* genes, small amplicons (300-600bp) encompassing the residues of interest were amplified by PCR as described above with the following primers:

GRN_W449_F1	TGTGAGGACAAGGAGCACTG
GRN_W449_R1	CAAACTCCATGCAGAAAGAGC
GRN_W449_F2	CACTTACTGAAACTTCCTCCACTGT
GRN_W449_R2	GCTGCTAAACAATGAAATATTCCTG
GRN_Q151_F1	TCATTCCAGAGTTGATTTTCACA
GRN_Q151_R1	AAGGGCAGACGTTGTGTACC
IFI35_M196_F	CATCTCATTAGTGGCGAGCA
IFI35_M196_R	AGAGTCGATCTGTGGGATGG
ZNF800A_N489_F	CCGCTGTTAGACTCCTCGTC
ZNF800A_N489_R	CAGAGTGTCGCCATGAAAGA

Genotype of individuals was determined by direct Sanger sequencing of PCR products (MCLAB sequencing services). Sequences were manually confirmed using the chromatogram profiles to detect variants at the homozygous or heterozygous state (Table S7E, F).

Supplemental References

Baumgart, M., Groth, M., Priebe, S., Savino, A., Testa, G., Dix, A., Ripa, R., Spallotta, F., Gaetano, C., Ori, M., et al. (2014). RNA-seq of the aging brain in the short-lived fish *N. furzeri* - conserved pathways and novel genes associated with neurogenesis. Aging Cell *13*, 965-974.

Capriotti, E., Calabrese, R., and Casadio, R. (2006). Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics *22*, 2729-2734.

Catchen, J.M., Amores, A., Hohenlohe, P., Cresko, W., and Postlethwait, J.H. (2011). Stacks: building and genotyping Loci de novo from short-read sequences. G3 *1*, 171-182.

Charif, D., and Lobry, J.R. (2007). SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In Structural Approaches to Sequence Evolution: Molecules, Networks, Populations (Biological and Medical Physics, Biomedical Engineering), U. Bastolla,

M. Porto, H.E. Roman, and M. Vendruscolo, eds. (Berlin, Heidelberg, Germany: Springer Verlag), pp. 207–232.

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) *6*, 80-92.

Cruts, M., Theuns, J., and Van Broeckhoven, C. (2012). Locus-specific mutation databases for neurodegenerative brain diseases. Hum. Mutat. *33*, 1340-1344.

Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., et al. (2014). Ensembl 2015. Nucleic Acids Res. *43*, D662-669.

Darriba, D., Taboada, G.L., Doallo, R., and Posada, D. (2011). ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics *27*, 1164-1165.

de Magalhaes, J.P., and Toussaint, O. (2004). GenAge: a genomic and proteomic network map of human ageing. FEBS Lett. *571*, 243-247.

Dehouck, Y., Kwasigroch, J.M., Gilis, D., and Rooman, M. (2011). PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. BMC Bioinformatics *12*, 151.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15-21.

Earl, D., Bradnam, K., St John, J., Darling, A., Lin, D., Fass, J., Yu, H.O., Buffalo, V., Zerbino, D.R., Diekhans, M., et al. (2011). Assemblathon 1: a competitive assessment of de novo short read assembly methods. Genome Res. *21*, 2224-2241.

Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics *26*, 2460-2461.

Etter, P.D., and Johnson, E. (2012). RAD paired-end sequencing for local de novo assembly and SNP discovery in non-model organisms. Methods Mol. Biol. *888*, 135-151.

Falcon, S., and Gentleman, R. (2007). Using GOstats to test gene lists for GO term association. Bioinformatics *23*, 257-258.

Fletcher, W., and Yang, Z. (2010). The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. Mol. Biol. Evol. *27*, 2257-2267.

Frappier, V., Chartier, M., and Najmanovich, R.J. (2015). ENCoM server: exploring protein conformational space and the effect of mutations on protein function and stability. Nucleic Acids Res. *43*, W395-400.

Gems, D., Sutton, A.J., Sundermeyer, M.L., Albert, P.S., King, K.V., Edgley, M.L., Larsen, P.L., and Riddle, D.L. (1998). Two pleiotropic classes of *daf-2* mutation affect larval arrest, adult behavior, reproduction and longevity in *Caenorhabditis elegans*. Genetics *150*, 129-155.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. *59*, 307-321.

Harel, I., Benayoun, B.A., Machado, B., Singh, P.P., Hu, C.K., Pech, M.F., Valenzano, D.R., Zhang, E., Sharp, S.C., Artandi, S.E., et al. (2015). A platform for rapid exploration of aging and diseases in a naturally short-lived vertebrate. Cell *160*, 1013-1026.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell *38*, 576-589.

Jordan, G., and Goldman, N. (2012). The effects of alignment error and alignment filtering on the sitewise detection of positive selection. Mol. Biol. Evol. *29*, 1125-1139.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. *110*, 462-467.

Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772-780.

Kent, W.J. (2002). BLAT - the BLAST-like alignment tool. Genome Res. *12*, 656-664.

Kim, D., and Salzberg, S.L. (2011). TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. Genome Biol. *12*, R72.

Kinsella, R.J., Kahari, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., Almeida-King, J., Staines, D., Derwent, P., Kerhornou, A., et al. (2011). Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database (Oxford) *2011*, bar030.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome Res. *19*, 1639-1645.

Lagesen, K., Hallin, P., Rodland, E.A., Staerfeldt, H.H., Rognes, T., and Ussery, D.W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. *35*, 3100-3108.

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., Huang, Q., Cai, Q., Li, B., Bai, Y., et al. (2010). The sequence and de novo assembly of the giant panda genome. Nature *463*, 311-317.

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658-1659.

Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923-930.

Liu, X., Han, S., Wang, Z., Gelernter, J., and Yang, B.Z. (2013). Variant callers for next-generation sequencing data: a comparison study. PLoS One *8*, e75619.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Lowe, T.M., and Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. *25*, 955-964.

Magoc, T., and Salzberg, S.L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics *27*, 2957-2963.

Marcais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics *27*, 764-770.

Marchler-Bauer, A., Derbyshire, M.K., Gonzales, N.R., Lu, S., Chitsaz, F., Geer, L.Y., Geer, R.C., He, J., Gwadz, M., Hurwitz, D.I., et al. (2015). CDD: NCBI's conserved domain database. Nucleic Acids Res. *43*, D222-226.

Michaelson, J.J., Alberts, R., Schughart, K., and Beyer, A. (2010). Data-driven assessment of eQTL mapping methods. BMC Genomics *11*, 502.

Mullan, L.J., and Bleasby, A.J. (2002). Short EMBOSS User Guide. European Molecular Biology Open Software Suite. Brief. Bioinform. *3*, 92-94.

Nawrocki, E.P., and Eddy, S.R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics *29*, 2933-2935.

Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics *23*, 1061-1067.

Parthiban, V., Gromiha, M.M., and Schomburg, D. (2006). CUPSAT: prediction of protein stability upon point mutations. Nucleic Acids Res. *34*, W239-242.

Patel, D.S., Garza-Garcia, A., Nanji, M., McElwee, J.J., Ackerman, D., Driscoll, P.C., and Gems, D. (2008). Clustering of genetically defined allele classes in the *Caenorhabditis elegans* DAF-2 insulin/IGF-1 receptor. Genetics *178*, 931-946.

Peel, M.C., Finlayson, B.L., and McMahon, T.A. (2007). Updated world map of the Köppen-Geiger climate classification. Hydrol. Earth Syst. Sci. *11*, 1633-1644.

Pei, J., Kim, B.H., and Grishin, N.V. (2008). PROMALS3D: a tool for multiple protein sequence and structure alignments. Nucleic Acids Res. *36*, 2295-2300.

Petzold, A., Reichwald, K., Groth, M., Taudien, S., Hartmann, N., Priebe, S., Shagin, D., Englert, C., and Platzer, M. (2013). The transcript catalogue of the short-lived fish *Nothobranchius furzeri* provides insights into age-dependent changes of mRNA levels. BMC Genomics *14*, 185.

Pires, D.E., Ascher, D.B., and Blundell, T.L. (2014). DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic Acids Res. *42*, W314-319.

Rembold, M., Lahiri, K., Foulkes, N.S., and Wittbrodt, J. (2006). Transgenesis in fish: efficient selection of transgenic fish by co-injection with a fluorescent reporter construct. Nat. Protoc. *1*, 1133-1139.

Salzberg, S.L., Phillippy, A.M., Zimin, A., Puiu, D., Magoc, T., Koren, S., Treangen, T.J., Schatz, M.C., Delcher, A.L., Roberts, M., et al. (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome Res. *22*, 557-567.

Schulz, M.H., Zerbino, D.R., Vingron, M., and Birney, E. (2012). Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics *28*, 1086-1092.

Sebastiani, P., Solovieff, N., Dewan, A.T., Walsh, K.M., Puca, A., Hartley, S.W., Melista, E., Andersen, S., Dworkis, D.A., Wilk, J.B., et al. (2012). Genetic signatures of exceptional longevity in humans. PLoS One *7*, e29848.

Simpson, J.T. (2014). Exploring genome characteristics and sequence quality without a reference. Bioinformatics *30*, 1228-1235.

Slater, G.S., and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics *6*, 31.

Smit, A.F.A., Hubley, R., and Green, P. (1996-2004). RepeatMasker Open-3.0. http://wwwrepeatmaskerorg.

Talavera, G., and Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. *56*, 564-577.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol. Biol. Evol. *30*, 2725-2729.

Tingaud-Sequeira, A., Lozano, J.J., Zapater, C., Otero, D., Kube, M., Reinhardt, R., and Cerda, J. (2013). A rapid transcriptome response is associated with desiccation resistance in aerially-exposed killifish embryos. PLoS One *8*, e64410.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. *7*, 562-578.

Treangen, T.J., and Salzberg, S.L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat. Rev. Genet. *13*, 36-46.

Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M., and Barton, G.J. (2009). Jalview Version 2 - a multiple sequence alignment editor and analysis workbench. Bioinformatics *25*, 1189-1191.

Wittkop, T., Emig, D., Lange, S., Rahmann, S., Albrecht, M., Morris, J.H., Bocker, S., Stoye, J., and Baumbach, J. (2010). Partitioning biological data with transitivity clustering. Nat. Methods *7*, 419-420.

Wong, W.S., Yang, Z., Goldman, N., and Nielsen, R. (2004). Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics *168*, 1041-1051.

Yang, Z., and dos Reis, M. (2011). Statistical properties of the branch-site test of positive selection. Mol. Biol. Evol. *28*, 1217-1228.

Yang, Z., and Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol. Biol. Evol. *19*, 908-917.

Yang, Z., Wong, W.S., and Nielsen, R. (2005). Bayes empirical bayes inference of amino acid sites under positive selection. Mol. Biol. Evol. *22*, 1107-1118.

Yao, G., Ye, L., Gao, H., Minx, P., Warren, W.C., and Weinstock, G.M. (2012). Graph accordance of next-generation sequence assemblies. Bioinformatics *28*, 13-16.